

<div style="display: flex; justify-content: space-between;"> REPORT SECURITY CLASSIFICATION D DD FORM DOCUMENTATION PAGE </div>											
AD-A211 630		ECTE 21 1989 B	D								
		15 ABSTRACTIVE MARKINGS <div style="border: 1px solid black; padding: 2px; display: inline-block; font-weight: bold;">FILE COPY</div>									
		1 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.									
		3 MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 89-1090									
6a. NAME OF PERFORMING ORGANIZATION	6b. OFFICE SYMBOL (if applicable)	7a. NAME OF MONITORING ORGANIZATION									
University of Minnesota		Air Force Office of Scientific Research/XL									
6c. ADDRESS (City, State, and ZIP Code)		7b. ADDRESS (City, State, and ZIP Code)									
1919 University Avenue St. Paul, MN 55104		Building 410 Bolling AFB, DC 20332-6448									
8a. NAME OF FUNDING/SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (if applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER									
AFOSR	NL	AFOSR-89-0232									
8c. ADDRESS (City, State, and ZIP Code)		10 SOURCE OF FUNDING NUMBERS									
Building 410 Bolling AFB, DC 20332-6448		<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <th style="text-align: left;">PROGRAM ELEMENT NO.</th> <th style="text-align: left;">PROJECT NO.</th> <th style="text-align: left;">TASK NO.</th> <th style="text-align: left;">WORK UNIT ACCESSION NO.</th> </tr> <tr> <td>61102F</td> <td>2313</td> <td>A5</td> <td></td> </tr> </table>		PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.	61102F	2313	A5	
PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.								
61102F	2313	A5									
11. TITLE (Include Security Classification)											
A Conference on Three-Dimensional Representation											
12. PERSONAL AUTHOR(S)											
Professor Irving Biederman											
13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT								
Final	FROM 01 Jan 89 TO 31 Dec 89	June 89	50								
16. SUPPLEMENTARY NOTATION											
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)									
FIELD	GROUP			SUB-GROUP							
19. ABSTRACT (Continue on reverse if necessary and identify by block number)											
<p>→ This is the final report for a conference grant entitled: "A Conference on Three-Dimensional Representation". The two and one-half day conference was held at the University of Minn. on May 24-26, 1989 to evaluate the current status of problems associated with three-dimensional representations from current computational, psychological, development, and neuro-physiological perspectives. Nineteen presentations were made spanning these approaches. One hundred sixty-six individuals attended the conference. Of 44 evaluations received, 75% rated the conference as Excellent, 20% as good, and 5% as fair. None rated it poor. The report consists of the original and revised program, conference abstracts evaluation summary and the roster of attendees.</p> <p><i>Keywords: depth perception, binocular space perception, three dimensional vision, object recognition, spatial processing, depth cues, scene, (KIR) ←</i></p>											
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT		21. ABSTRACT SECURITY CLASSIFICATION									
<input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS		UNCLASSIFIED									
22a. NAME OF RESPONSIBLE INDIVIDUAL		22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL								
Dr. John F. Tannoy		(202) 767-3021	NL								

DD FORM 1473, 84 MAR

 8) APR edition may be used until exhausted.
 All other editions are obsolete.

 SECURITY CLASSIFICATION OF THIS PAGE
 UNCLASSIFIED

89 8 18 049

FINAL REPORT

Submitted to: Air Force Office of Scientific Research
AFOSR/NL
Building 410
Bolling A.F.B., D.C. 20332-6448

Dr. John Tangney, Program Manager

Title: A CONFERENCE ON THREE
DIMENSIONAL REPRESENTATION


AFOSR Grant No. AFOSR-89-0232

Report Security
Classification: Unclassified

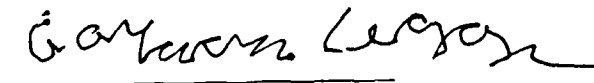
Duration: 1 JAN 89-31 DEC 89

Submitting
Organization: University of Minnesota
Office of Research Administration
1919 University Ave.
St. Paul, MN 55104

PIs: Professor Irving Biederman
(612) 626-0807


Irving Biederman
134-28-9789

Professor Gordon E. Legge
(612) 625-0806


Gordon E. Legge
023-42-1595

Department of Psychology
University of Minnesota
Elliott Hall
75 East River Road
Minneapolis, MN 55455

FINAL REPORT

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH GRANT 89-0232

A CONFERENCE ON THREE-DIMENSIONAL REPRESENTATION

Abstract

This is the final report for a conference grant entitled "A Conference on Three-Dimensional Representation." The two and one-half day conference was held at the University of Minnesota on May 24-26, 1989 to evaluate the current status of problems associated with three-dimensional representations from current computational, psychological, developmental, and neurophysiological perspectives. Nineteen presentations were made spanning these approaches. One hundred sixty-six individuals attended the conference. Of 44 evaluations received, 75 percent rated the conference as Excellent, 20 percent as good, and 5 percent as fair. None rated it poor.

The report consists of the original and revised program, conference abstracts, evaluation summary, and the roster of attendees.

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	



VISION AND THREE-DIMENSIONAL REPRESENTATION

May 24-26, 1989 Electrical Engineering and Computer Science Building University of Minnesota Minneapolis, Minnesota

FINAL PROGRAM

WEDNESDAY, MAY 24

Electrical Engineering/Computer Science Building,
Room 3-210

7:30 a.m.

Registration and Refreshments

8:00

Opening Remarks

Gordon Legge, Conference Co-Chair, Department of
Psychology, University of Minnesota

Session I

Chair: Gordon Legge, Department of Psychology

8:30

Development of Depth Perception

Albert Yonas, Institute of Child Development,
University of Minnesota

9:20

Binocular Space Perception

John M. Foley, Department of Psychology, University
of California, Santa Barbara

10:10

Break

Session II

*Chair: Jonathan Marshall, Minnesota Supercomputer
Institute and Center for Research in Learning, Perception,
and Cognition*

10:30

The Reconstruction of Continuous Surfaces from
Stereo Measurements and Monocular Inferences

Kent A. Stevens, Department of Computer Science,
University of Oregon

11:20

Occlusion Constraints and the Encoding of Color,
Form, Motion, and Depth

Ken Nakayama, Smith-Kettlewell Eye Institute

12:10 p.m.

Lunch (on your own)

Session III

Chair: Lincoln Craton, Institute of Child Development

2:00

How 3-D Are We?

Patrick Cavanagh, Department of Psychology,
University of Montreal

2:50

Analyzing Visual Motion—Spatial Organization at
Surface Boundaries

William B. Thompson, Department of Computer Science,
University of Minnesota

3:40

Break

4:00

Perception of 3-D Structure from Motion

James Todd, Department of Psychology, Brandeis
University

5:00

Dinner (on your own)

Session IV

Session Chair: J.R. Barnes, General College

8:00

A Perceptual Approach to Art: Comments on the Art
Exhibition, Mississippi Room, third floor, Coffman
Memorial Union

Bruce Goldstein, Department of Psychology, University
of Pittsburgh

9:00

Reception and viewing of the art exhibition in the
Coffman Gallery, first floor

THURSDAY, MAY 25

8:00 a.m.

Refreshments

Session V

*Chair: Lee Zimmerman, Department of Electrical
Engineering*

8:30

Representing Constraints for Inferring 3-D Scene
Structure from Monocular Cues

Jitendra Malik, Department of Computer Science,
University of California, Berkeley

9:20

Viewpoint Invariant Primitives as a Basis for Human Object Recognition

Irving Biederman, Department of Psychology,
University of Minnesota

10:10

Break

Session VI

Chair: Liz Stuck, Department of Computer Science

10:30

Representation and the Scene Modeling Problem

Martin Fischler, SRI International

11:20

Object Recognition from Range Images

Ramesh Jain, Department of Electrical Engineering and
Computer Science, University of Michigan

12:10 p.m.

Lunch (on your own)

Session VII

Chair: David Parish, Department of Psychology

1:30

Perception and Perceivers

Allan Jepson, University of Toronto
Whitman Richards, Massachusetts Institute of
Technology

2:20

**Parallel Processing of Form, Color, Motion and Depth
in the Visual System**

Margaret Livingstone, Department of Neurophysiology,
Harvard Medical School

3:10

Break

Session VIII

Chair: John Hilton, Department of Psychology

3:30

**Cortical Pathways for the Analysis of Form, Space, and
Motion: Three Streams of Visual Processing**

Leslie G. Ungerleider, NIMH Laboratory of
Neuropsychology

4:20

Components of High-Level Vision

Stephen M. Kosslyn, Department of Psychology,
Harvard University

7:00

Barbecue at the home of Dr. Irving Biederman,
6612 Cornelia, Edina.

FRIDAY, MAY 26

8:00 a.m.

Refreshments

Session IX

Chair: Martha Arterberry, Institute of Child Development

8:30

**Depth Cues and the Computation of Real Distance:
The Calibration of Ballistic Movements**

Melvin A. Goodale, Department of Psychology,
University of Western Ontario

9:20

**Constraints Imposed by Occlusion and Image-
Segmentation**

V.S. Ramachandran, Department of Psychology,
University of California, San Diego

10:10

Break

10:30

**Binocular Representation of the Visual Field in
Primate Cortex**

Eric L. Schwartz, New York University Medical Center,
Brain Research Laboratory

11:20

Closing remarks

Lee Zimmerman, Conference Co-Chair, Department of
Electrical Engineering, University of Minnesota

11:30

Conclusion

ABSTRACTS

Vision and Three-Dimensional Representation

May 24-26, 1989

University of Minnesota

Sponsored by
Air Force Office of Scientific Research
College of Liberal Arts, University of Minnesota
Center for Research in Learning, Perception, and Cognition

These compiled abstracts are from presentations given at the conference entitled "Vision and Three-Dimensional Representation" held May 24-26, 1989 at the University of Minnesota.

CONFERENCE SPONSORS

Air Force Office of Scientific Research
College of Liberal Arts, University of Minnesota
Center for Research in Learning, Perception, and Cognition

IN COOPERATION WITH

Department of Computer Science
Department of Electrical Engineering
Department of Psychology
Department of Studio Arts
Department of Institute of Child Development
Department of Professional Development Conference Services, and
Continuing Education and Extension

CONFERENCE CO-CHAIRS

Gordon Legge
Lee Zimmerman

CONFERENCE PLANNING COMMITTEE

Martha Arterberry, J.R. Barnes, Irving Biederman, Lisa Brienzo, Victor Caglioti, Lincoln Craton, Steven Cusulos, Jon Gottesman, Char Greenwald, John Hilton, Jonathan Marshall, Jo Nichols, Bruce Overmier, David Parish and Liz Stuck

FOR MORE INFORMATION CONTACT

Center for Research in Learning, Perception, and Cognition
University of Minnesota
203 Elliot Hall, 75 East River Road
Minneapolis, Minnesota 55455

ABSTRACTS AND AUTHORS
(in order of presentation)

- | | |
|--------------------------------------|---|
| 1. Albert Yonas | Development of Depth Perception |
| 2. John M. Foley | Binocular Space Perception |
| 3. Kent A. Stevens | The Reconstruction of Continuous Surfaces from Stereo Measurements and Monocular Inferences |
| 4. Ken Nakayama | Occlusion Constraints and the Encoding of Color, Form, Motion, and Depth |
| 5. Patrick Cavanagh | How 3-D Are We? |
| 6. William B. Thompson | Analyzing Visual Motion--Spatial Organization at Surface Boundaries |
| 7. James Todd | Perception of 3-D Structure from Motion |
| 8. Jittendra Malik | Representing Constraints for Inferring 3-D Scene Structure from Monocular Views |
| 9. Irving Biederman | Viewpoint Invariant Primitives as a Basis for Visual Object Recognition |
| 10. Martin A. Fischler | Representation and the Scene Modeling Problem |
| 11. Ramesh Jain | Object Recognition from Range Images |
| 12. Allan Jepson
Whitman Richards | Perception and Perceivers |
| 13. Margaret Livingstone | Parallel Processing of Form, Color, Motion, and Depth in the Visual System |
| 14. Leslie G. Ungerleider | Cortical Pathways for the Analysis of Form, Space, and Motion: Three Streams of Visual Processing |
| 15. Stephen M. Kosslyn | Components of High-Level Vision |
| 16. Melvin A. Goodale | Depth Cues and the Computation of Real Distance: The Calibration of Ballistic Movements |
| 17. V.S. Ramachandran | Constraints Imposed by Occlusion and Image-Segmentation |
| 18. Eric L. Schwartz | Binocular Representation of the Visual Field in Primate Cortex |

Infants' Sensitivity to Time-Dependent Information
Albert Yonas and Martha E. Arterberry
University of Minnesota

The traditional approach to understanding a complex process or phenomenon is to divide and conquer. Find the right parts of a process and work to explain the parts. We have faith that the parts will add up to the whole. This approach has been remarkably successful in the physical and biological sciences. It seems fair to say that in the study of cognition we have found it difficult to describe the proper parts of the process by which we gain information about the world. For example there has been a long history of argument over whether there is a single process or two distinct macro processes that intervene between the detection of light that strikes the retina and the result, i.e. what is known about the world by the viewer. Those two processes have been termed perception and cognition.

James Gibson argued for a single perceptual process in which there is direct pickup of higher-order amodal invariants that specify information that allows us to effectively control our actions and to obtain veridical knowledge of the environment. Helmholtz and his followers, such as Richard Gregory and Irvin Rock, also argued that perception and cognition could best be accounted for by a single set of processes. But unlike Gibson, who wanted to account for cognition by pointing out its similarity to sensory functioning, Helmholtz argued that perception was the result of inference-like processes similar to those that occurred in thinking and problem solving, i.e. the theory of unconscious inference.

It is our claim that through a comparative approach, which explores the development of behavior across species and over time, we can better cut nature at its joints and explore the fundamental elements of cognition.

Over the last twenty years we have asked questions about the development of four abilities in human infants:

1. When does the ability to build a representation of a spatial layout (particularly, object length and number) develop when that layout cannot be perceived at any point in time, but must be perceived over time?

2. How does the ability to use pictorial depth cues develop? When does perception of spatial layout appear if information does not have to be integrated over time and only static monocular depth

cues are available? 3. When does the ability to use binocular depth information develop?

4. When does the ability to use kinetic depth information develop?

The take home message of this paper is that our work suggests these four abilities are not accounted for by a single process that develops at a single time. Rather, in the human infant, these abilities are demonstrated at four different times over the first year of life.

1. Recent studies by Martha Arterberry and myself suggest that the ability to use information that is only available over time to build a mental representation of object length and number appears between the 8th and 12th months. (In these studies objects are viewed as they are moved behind a window so that at any point in time one can not discriminate, for example, between a display of a long object or a short object).

2. A second series of studies carried out by Carl Granrud, myself and others suggests that infants begin to perceive spatial layout from information provided by pictorial depth cues between the 5th and 7th months.

3. A third body of literature, contributed to by Held, Aslin and many others, has demonstrated that sensitivity to binocular information normally develops in the 4th month in human infants.

A final set of studies carried out by Bennett Bertenthal at Virginia, Phillip Kellman at Swarthmore, by our group at Minnesota and by others suggests that sensitivity to kinetic information may develop before sensitivity to binocular information. Sensitivity to optical information for collision, "looming" appears by the end of the first month and may be present at birth. Lincoln Craton and I have recently found evidence that 3 and 1/2 month olds can segment figure from ground using purely kinetic cues (boundary flow and accretion and deletion of texture). At this time we know little about the structure of the processes that compute kinetic information. There may not be a single box that accounts for kinetic sensitivity. Rather, there may obtain a bag of independent "tricks" that develop, at different times, that detect different cues. In any event, it seems clear that long before the infant uses static monocular information, he can use kinetic spatial information.

We would like to argue that developmental findings like these inform us about the structure of the mechanisms that make it possible for the infant to develop into a child who, at 12 months, perceives the environment with remarkable effectiveness and controls his action with amazing skill.

BINOCULAR SPACE PERCEPTION

John M. Foley

Having two eyes with a space between them adds considerably to an organism's potential for space perception. If the direction of an object from each eye and the distance between the eyes are known, a simple calculation yields the distance of the object from the eyes. Since light carries information about the directions of objects, light arriving at the two eyes from the same object carries information about distance as well. This talk is concerned with the accuracy with which human observers respond on the basis of binocular information to the locations of objects and the extents between them.

It will be shown that there are large and systematic errors in binocular distance perception and that the binocular visual space is grossly distorted in relation to physical space. Its intrinsic geometry differs greatly from Euclidean geometry. A formal model of binocular space perception will be described and the processes underlying this model will be considered.

The tasks appropriate to the study of space perception are open loop tasks (tasks without feedback) in which an observer indicates the relative or absolute locations of objects (here called targets) or the distances between them. Binocular space perception refers to situations in which the stimulus variables that carry spatial information (cues) are present only in binocular vision, although it is difficult to design situations to answer some questions about binocular vision that exclude monocular cues.

The tasks of concern here are tasks in which the observer is asked to respond on the basis of perceived locations or extents. A distinction is made between relative tasks and absolute tasks. In a relative task one visible target or extent is judged in relation to another. Usually, they are present simultaneously, although they need not be. In an absolute task a target location or an extent between targets is compared with an internal representation of the space. For example, an observer reports that a target is perceived to be 10 deg to the right of straight ahead and 2 meters away. There are other tasks in which visual percepts are indicated by motor responses (e.g., pointing with an unseen hand). These responses cannot be taken as measurements of visual percepts, but may be related to them by constant transforms.

When human observers indicate the locations of visual targets and the magnitudes of visual extents, systematic errors are made in both absolute and relative tasks. Different indicator responses show different errors that are systematically related. The occurrence of such errors does not necessarily mean that the percept is in error; the errors might arise in processes that intervene between the percept and the response. Here it will be argued that the percept is in error. Transformations that intervene between the percept and the response modify this error and account for the different magnitudes of error that occur with different types of indicator responses.

The binocular perception of direction and distance will first be considered. It will be shown that the data from both relative and absolute tasks are consistent with a model in which perceived distance depends on the integration of an egocentric distance signal and a disparity signal. The first is systematically in error and the second is accurate (under some conditions). Evidence relating to the geometry of visual space and its associated metric will then be described. Neither a Euclidean nor a Riemannian metric describes the binocular perception of extent. Another metric is proposed. When binocular images contain many identical closely-spaced points, as in random-dot stereograms, there is ambiguity as to which points correspond. Several models have been proposed of how disparity is computed for such images. These generally predict the accurate computation of disparity in most instances, a prediction that is consistent with the model described above. However, evidence from experiments with complex stereograms indicates that for such stimuli disparity is often not accurately determined by the visual system. Thus, the general relation between disparity and its representation in the visual system (effective disparity) remains to be determined. Finally, possible processes underlying the distortions in binocular space perception will be considered.

The Reconstruction of Continuous Surfaces from Stereo Measurements and Monocular Inferences

Kent A. Stevens
Department of Computer Science
University of Oregon

How does one reveal the nature of the visual representation of surfaces? Informally, this representation is expected to buffer or store three-dimensional information about perceived surfaces derived from various visual cues. The representation is expected to subserve, in natural vision, such perceptual tasks as object recognition and manipulation, and in the laboratory, simpler tasks such as judging the orientation of a depicted surface, or the relative depth of two points in an image. It takes a modest leap of faith to see a connection between such artificial judgments and the natural processes that might need such information.

Analysis of what specific information might theoretically be delivered by specific cues has led to the general expectation that several three-dimensional quantities are made explicit simultaneously. Local quantities such as the orientation, curvature, and distance of local surface patches (usually proposed to be relative to the observer) are often expected to be represented straightforwardly by two-dimensional arrays indexed by visual direction. Arrays (or maps) might be used to represent spatially extended surface features, such as loci where the surface is sharply discontinuous or creased, or topographic features such as ridges. Multi-resolution schemes extend these ideas to account for the fact that a complex surface may have multiple descriptions, based on the spatial scale at which the descriptors are applied.

Empirically demonstrating the relevance of these various theoretical proposals to human vision has proven surprisingly difficult. Some seemingly conflicting observations have been reported, which probably stem from the classical Gestalt notion of *Praeganz*, that the percept will be as good as the stimulus allows. In cue reduction experiments, where only limited three-dimensional information is presented (in sharp contrast to most natural images) the three-dimensional judgments (particularly of slant) might be highly unreliable, while at the same time more qualitative judgments such as the ordinal depth relationship of two surface points remain more stable. Other reduction experiments, however, using different but still highly impoverished stimuli, have shown that observers can make stable and accurate slant judgments, provided the stimulus provides sufficient information to suggest a surface of definite orientation. Since the quality of the three-dimensional judgment depends critically on the quality of the stimulus information, it is difficult to gain insight into the nature of the underlying representation by reducing the quality of the stimulus.

The psychological investigation of surface representation has often focussed on *what* information is stored (when not addressing the closely related question of *how* that information is extracted). The recent influx of computer models, particularly those using arrays as the medium for representing such information, has probably induced some false optimism in this pursuit, because in those demonstrations there is rarely much concern given to the problem of subsequently accessing the information stored in the array. The visual system does not merely store information in some passive map or array; it actively subserves many tasks of the organism. This point has long been obvious but has somehow been underemphasized of late. To help remedy the situation, it might be of some use to borrow from computer science a concept that focuses attention on the idea of data as being active rather than passive. Importantly, for the same reason it has been helpful to

computer science it will be helpful to the current pursuit. Namely, by defining so-called abstract data types (ADTs), one can be highly specific about what information might be stored, and about what information might be accessed, without concern for precisely how the information is implemented.

Describing information processing in terms of abstract data types allows one to make headway in specifying a system without distracting oneself with the implementation details (such as whether to represent local scalar information in a map, and if so, whether that map needs to be congruent and in spatial registration with another map storing related information). Questions of greater concern would be *what information is used to subserve what tasks*, and *what tasks have access to what information* (and perhaps exclusion from other information).

I will discuss recent work in surface perception, particularly in the resolution (or integration) of surface information from stereopsis and monocular sources, attempting to keep the focus on what information is effective and what is ineffective. The more I have worked in surface perception, the less certain I am about any aspects of the details of the representations. While in wholehearted agreement with recent arguments against point-by-point mappings of depth and of surface orientation (e.g. Todd & Reichel, in press), there is nonetheless good evidence that these properties can be provided with considerable quantitative precision, if the image supports it, as mentioned. Given our experience, we would tend to trust that a given type of information is represented if it could be shown to be effective in a specific perceptual task. Rather than trust the results of experiments involving direct psychophysical judgments, we have concentrated on experiments for which the judgments would reflect only the indirect contribution of the specific three-dimensional information in question. In examining the integration of three-dimensional information from stereopsis and monocular sources we have found certain information long expected to be represented (such as pointwise depth from stereopsis) to be remarkably ineffective towards the final percept, under certain controllable circumstances. In terms of the ADT metaphor, we are addressing representation problems in terms of what processes have access to what sorts of information. The discussion will draw on the following citations.

Stevens, K.A. 1981 The visual interpretation of surface contours. *Artificial Intelligence* 217, Special Issue on Computer Vision, 47-74.

Stevens, K.A. 1981 The information content of texture gradients. *Biological Cybernetics* 42, 95-105.

Stevens, K.A. 1983 Surface tilt (the direction of surface slant): a neglected psychophysical variable. *Perception and Psychophysics* 33, 241-250.

Stevens, K.A. 1983 The line of curvature constraint and the interpretation of 3-D shape from parallel surface contours. *Eighth International Joint Conference on Artificial Intelligence*, August.

Stevens, K.A. 1983 Slant-tilt: The visual encoding of surface orientation. *Biological Cybernetics* 46, 183-195.

Stevens, K.A. 1984 On gradients and texture "gradients". Commentary on: Cutting & Millard 1984. Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General* 113, 217-220.

Stevens, K.A. 1986 3-D shape from 2-D contour. Invited paper, *Proceedings of the Annual Meeting of the Optical Society of America*, Seattle, Washington, October.

Stevens, K.A. 1986 Inferring shape from contours across surfaces. In *From Pixels to Predicates: Recent Advances in Computational Vision*. A.P. Pentland, ed., 93-110. Norwood, N.J.: Ablex.

Stevens, K.A. & Brookes, A. 1987 Probing depth in monocular images. *Biological Cybernetics* 56, 355-366.

Stevens, K.A. & Brookes, A. 1987 Depth reconstruction in stereopsis. *Proceeding of the First IEEE International Conference on Computer Vision*, London, June.

Stevens, K.A. & Brookes, A. 1988 Integrating stereopsis with monocular interpretations of planar surfaces. *Vision Research* 28, 371-386.

Brookes, A. & Stevens, K.A. 1988 Binocular depth from surfaces vs. volumes. *Journal of Experimental Psychology: Human Perception and Performance*, in press.

Brookes, A. & Stevens, K.A. The analogy between stereo depth and brightness. *Perception*, in press.

Todd, J.T. and Reichel, F.D. Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychological Review*, in press.

Occlusion constraints and the encoding of color, form, motion and depth

Ken Nakayama

Smith Kettlewell Eye Research Institute

San Francisco, CA

ABSTRACT

We live in a three dimensional world, full of non-transparent objects and surfaces. As a consequence, objects occlude other objects, and the boundaries and surfaces of occluded objects are often only partially visible. Thus, occlusion is one of the most fundamental facts about vision in daily life and is decisive in determining how light from the physical world reaches our eyes. Consequently, it sets a major obstacle which the visual system must overcome to accomplish its goal of identifying 3-dimensional objects from purely viewer centered visual inputs.

Recent research in our laboratory conducted in collaboration with Dr. Shinsuke Shimojo indicates that the encoding of color, form, motion, and depth is highly influenced by such occlusion constraints. Our work can be divided into three main areas: (1) the intrinsic/extrinsic classification of surface boundaries, (2) the encoding of binocularly unpaired image points, (3) the perception of transparency.

1. We hypothesize that bounding contours can be divided into two categories, those which are intrinsic to the surface itself and which provide useful information regarding its shape and motion, and those which

are extrinsic to this surface being only accidentally related via occlusion and which contain no information regarding the occluded surface. Our experiments suggest that the visual system treats these two types of boundary contours differentially, both for the purposes of pattern recognition and for the ambiguity solving process posed by the aperture problem. Furthermore, we suggest that the classification is based on a prior encoding of local depth relations.

2. If we consider binocular vision in an opaque world of surfaces, it is important to recognize that many points in the environment are visible only to one eye or the other. Thus background points immediately to the right of an occluding object will be seen by the right eye only and background points to the left of this object will be seen by the left eye only. We have discovered a set of perceptual phenomena which indicates that the brain handles these unpaired points in a remarkably adaptive manner, seeming to take account of the real world scene geometry to get useful information about depth and surface properties. In particular, we have found that: a) isolated monocular-only points give rise to subjective contours which are localized immediately to the right of left-eye-only points and immediately to the left of right-eye-only points. b) monocular points presented in ecologically valid configurations will be seen as non-rivalrous and generally in back of adjacent fused binocular points. c). horizontal movement presented sequentially to the two eyes with no temporal overlap can appear as a single target moving in depth if it simulates the ecologically valid situation of a target moving behind a narrow slit.

Since information regarding the eye-of-origin information of these monocular-only points is decisive for these perceptual effects and since

this information is presumably lost by V2 (Burkhalter and van Essen, 1986), we suggest that the substrate for the functional analysis of these occluded image points may begin at early stages of cortical processing.

3. Finally, we have considered the phenomenon of transparency, the converse of opacity. Using stereoscopic techniques we have shown that the perception of transparency is dependent on the more primitive features of contour, luminance, and depth. Yet, in turn, the perception of transparency itself can have a major role in the encoding of depth, contour and color (as in neon-color spreading).

How 3-D Are We?

Patrick Cavanagh
Département de Psychologie
Université de Montréal
Montréal, Québec
Canada H3C 3J7

We have begun to study distortions of rigid, 3-D objects viewed binocularly. These include folded drawings, wire frame objects and solid objects. The distortions may be explained by an inappropriate weighting of binocular information in forming the internal 3-D representation but interestingly, the weighting for a given stimulus appears to depend on the task. These studies, along with others on shadow images, suggest that the visual system may begin image analysis with a 2-D matching process and use the results of this process to guide 3-D modeling. Matched filtering or template matching are examples of the type of 2-D matching implied by our results.

Although template matching has long been a favorite example of how not to do pattern recognition (to the extent that its current proponents have relabeled it "model-driven" object recognition), all approaches to visual recognition use templates at some level. The receptive fields that underlie contour identification are straightforward template (or convolution) operators. Therefore, the question is not whether to use templates, but at what level. Typically, the results of template operations are labeled at an early stage as particular image tokens — edges or curves — to be used in further analysis. This early labeling commits the visual analysis to treat image elements in specific ways in subsequent processing and it can be disadvantageous to make this commitment prematurely. Many image contours, for example, cannot be unambiguously identified until the object to which they belong has been identified. In the case of an image with shadow and material borders, it is often impossible to know from the image whether a border is in fact a shadow border or an object border. The segmentation of a continuous border that is a shadow border in one region but an object contour in another (a material change or external contour) is particularly difficult. Any

approach that labels the borders before identifying the object will be faced with several cast shadow borders that should, in fact, be discarded. These extra borders will seriously disrupt the segmentation of object parts into volumetric units — geons (Biederman, 1987, *Psychological Review* 84, 115-147) or generalized cones (Marr, 1982). This problem of parsing object contours before identifying the object underlines the advantages of an initial 2-D match between the image and memory prototypes of 2-D object views without first selecting which image contours must be object contours. General matched-filtering techniques, for example, allow a match to be determined based on partial images and, if a reversible transform is used, to highlight (in the image plane) those contours that participated in the match and to fill in the missing contours. The significant drawback of this technique is the formidable task of deciding which memory prototype produced the best match to which part of the image and getting rid of the competitors. This is a problem that cannot be solved by linear techniques and requires a cooperative labeling approach.

If initial memory contact is based on 2-D views, then each object would have to have several of its 2-D views stored as part of its representation. In one sense, this is not an insurmountable problem even in the worst case since a matched-filter approach is ideally suited to massively parallel memory access. Once memory access has identified the possible objects in the scene, stored 3-D information about the objects can then be used to construct the 3-D scene representation that receives the most image support.

There is, in fact, strong evidence that the visual system operates on viewer-centred (2-D) representations and not 3-D object models when accessing memory. Rock and his colleagues (Rock, DeVita & Barbeito, 1981, *JEP:HPP* 7, 719-731; Rock & DeVita, 1987, *Cognitive Psychology* 19, 280-293) have demonstrated that views of wire-frame objects seen from different directions are only reliably identified when they have the same retinal projection, indicating that 2-D viewer-centred representations may mediate recognition.

We have, as well, discovered a variety of shapes whose internal representation appears to depend on the task being performed. A rigid, wire-frame object appears to create a veridical 3-D

representation when a task of static 3-D localization is used, a nonrigid 3-D representation when the same shape is rotated in a structure-from-motion task, and essentially a 2-D representation (total loss of shape constancy) when the shape is used as the mouth of a schematic face and the judgement is one of facial expression. Our initial interpretation of these results is that several internal representations are created simultaneously and these are accessed in a manner that depends on the task requirements. The possibility of multiple internal representations of visual space has been previously suggested in other contexts (Goodale, Pelisson & Prablanc, 1986, *Nature* 320, 748-750).

ANALYZING VISUAL MOTION -- SPATIAL ORGANIZATION AT SURFACE BOUNDARIES

**William B. Thompson
Computer Science Department
University of Minnesota**

The motion of a sensor and/or objects under view produces distinctive changes over time in an image. Most contemporary research on interpreting visual motion concentrates on the estimation of precise object geometry based on subtle variation in temporal properties over object surfaces. In this talk, we show how visual patterns at surface boundaries also signal the spatial organization of a scene. We describe a more "qualitative" approach to visual analysis which is likely to be more general and reliable than the current computational models which depend on strict assumptions about the scene and require the solution to ill-conditioned systems of equations.

Motion-based boundary analysis is sensitive only to depth discontinuities and/or object boundaries. Thus, unlike edge detection based on static properties such as intensity, texture, or color, all detected edge points are of direct significance to the interpretation of object shape. Perhaps even more importantly, the patterns of change at a dynamic occlusion boundary usually allows the identification of occluding and occluded surfaces, providing important information about the three-dimensional spatial structure of the scene. We have been able to show that not only is this an effective technique for computer vision systems, but that it is also used as a depth cue by human vision. Finally, we describe the effects of eye movement on the motion-based perception of surface boundaries. A computational analysis is presented, along with a number of open questions about how biological vision systems might deal with this situation.

Perception of 3-D Structure from Motion

James Todd

Brandeis University

Most theoretical analyses of the visual perception of structure from motion have been primarily concerned with minimal conditions for computing the 3D metric structure of an object from its projected movements (or displacements) within a 2D visual image. One of the principal results in this area, first reported by Ullman (1979), is that the 3D structure of an arbitrary configuration cannot be uniquely determined from an apparent motion sequence, unless the sequence contains a minimum of three distinct images. During the past decade, numerous investigators have attempted to determine how closely this theoretical limit corresponds to the perceptual limitations of human observers, but it has proven to be especially difficult to develop appropriate methodological procedures to precisely measure the perceived 3D structure of an object with adequate controls to eliminate any possible information within each static image.

The research described in this paper was designed in an effort to address these issues using a variety of different response tasks and stimulus displays as converging operations. Two unexpected results have been obtained repeatedly in these experiments: First, observers are surprisingly poor at discriminating 3D metric structures (e.g., relative lengths or angles) if care is taken to eliminate static information from each individual image in the apparent motion sequence; and

second, although performance is significantly improved for 2-frame sequences over static controls, there are no further improvements as the number of frames is increased beyond two. Since current algorithms for computing structure from motion require a minimum of three distinct images, these findings indicate that existing theory may have little relevance to actual human perception.

One possible explanation of these results is that perceptual knowledge of moving objects in 3-space may be primarily nonmetric in nature. Nonmetric knowledge can involve categorical distinctions such as rigid vs. nonrigid, flat vs. curved, or smooth vs. rough. It can also be based on order relations (e.g., it is sometimes possible to determine that one point is closer in depth than another without knowing how much closer). To provide additional support for this hypothesis, evidence will be presented that judgments of nonmetric structure are significantly more accurate and have shorter reaction times than do comparable judgments of metric structure.

Representing constraints for inferring 3-D scene structure from monocular cues

Jitendra Malik
University of California, Berkeley

At least since Marr's $2\frac{1}{2}$ D sketch idea and Barrow and Tennenbaum's intrinsic images proposal, it has been widely believed in the computer vision community that the transformation from image descriptions to scene descriptions is performed by a set of shape-from-X modules. These modules e.g. shape from shading, contour, texture, stereo and motion operate virtually independently and result in the computation of pointwise depth and orientation.

If we apply the criterion—Do the algorithms work robustly on images of real scenes—we find that the work on the monocular shape-from-X modules has not been successful. This is because the mathematical models have dealt with micro-worlds. The definition of shape-from processes as separate modules inevitably leads to models where to isolate the effect of one factor, we have to assume that the other factors are known and constant. A prime example of this phenomenon is the study of shape from shading where in order to isolate the effect of shape, a simple, known reflectance map is assumed. We have handicapped ourselves in another way: by setting too ambitious a goal. Pointwise computation of depth from monocular cues is neither necessary (for object recognition) and perhaps impossible (for images of arbitrary scenes).

It is my opinion that shading, contour and texture constraints are closely coupled and partitioning their analysis into separate shape-from-X modules which interact only at the output stage is a bad idea. Here are some examples of the close coupling:

1. Shape-from-shading algorithms need boundary conditions which are available only after the contours in the line drawing have been labeled. The constraint at a limb is different from that at a tangent plane discontinuity edge.
2. In images like photographs of the crater illusion, rotating the image by 180° leads to the hollows being interpreted as bulges and vice versa. This is an instance of the shading gradient influencing the 3D interpretation of the bounding contours of the crater.

3. Line drawings, texture and shading are really prototypical representatives of a continuum. It would probably make sense to use one algorithm which adapts to the image, instead of three completely distinct algorithms.
4. Gilchrist *et al* have shown how even the perception of lightness—once believed to be a cleanly separable process—is influenced by the classification of edges as shadows, dihedral edges or reflectance discontinuities. This phenomenon had gone unnoticed in the Land style experiments because of the impoverished nature of the Mondrian stimuli used.

I believe that a major goal of form analysis is to organize the scene. By this we mean the (perhaps partial) classification of each image intensity discontinuity and the (perhaps partial) determination of coarsely quantized values of the surface orientation, albedo and depth. Object recognition is possible from these partial interpretations.

I will discuss how the representation of image and the scene structure as multiresolution Markov Random Fields is convenient for this task. At the coarsest resolution the image is decomposed into a set of regions determined by the line drawing; the MRF sites are junctions, curves or regions. At the finest resolution the sites are either (a) pixels and (b) those line sites which correspond to image intensity edges.

In the MRF framework constraints are modelled as clique potentials, and the problem of determining scene structure becomes one of minimizing the sum of all the clique potentials. Equivalently, we can regard this as a stochastic optimization problem with an objective function given by the sum of the clique potentials. Some examples of representing monocular constraints in this framework follow:

1. Surface smoothness constraints can be represented by terms of the type $\|n_i - n_j\|$ where i and j are neighboring pixels or groups of pixels at coarser levels.
2. Junction labeling constraints are represented by giving negative potentials to the cliques corresponding to the labellings in the 3-D junction catalog (Malik 87). By giving weights to the different terms we can allow for the fact that junctions may not have been classified

correctly. Note that giving preference to 3D interpretations has been achieved for free.

3. We give positive costs to the line label changing along an edge. This permits the line label to change if necessary but remain the same if possible. This solves the dilemma: if we allow arbitrary label changes there is a combinatorial explosion in number of labellings else one is forced to restrict convex to concave transitions by fiat e.g. in (Malik 87).
4. Surface shading constraints can be expressed in a natural way. The direction of the equivalent light source is a quasi-global variable. This is inferred as part of the interpretation process (default starting value: lighting is from above). Line labeling and surface interpolation (Barrow-Tennenbaum 81) generate crude orientation estimates which make this computation possible. (Brooks-Horn 85 demonstrate the plausibility of a related approach).

Note that MRFs are just one possible machinery for representing these constraints. The crucial idea here is that of representing the process of 3D scene recovery as a minimization problem. Any framework which permits the representation of the wide variety of monocular cues would be adequate for the purpose.

How are we to minimize this complicated looking function? Embedded as special cases are several NP-complete problems e.g. line labelling. It seems to us that there is no alternative to search. Search can be avoided only when we are sufficiently close to a global minimum and gradient descent will work. But the hard part, and perhaps the most essential part is that of getting close enough in the first place. Of course we can and should reduce search as far as possible. We feel that expressing constraints in a least commitment style, using very coarse quantizations both of regions and values of parameters like surface normals, and being clever about the choice of particular Monte Carlo optimization strategy makes the problem tractable.

And now for a speculative sketch of an algorithm. I believe that obtaining a single consistent interpretation is an inherently sequential process mediated by a Triesman type feature-integrating searchlight. Selecting an

interpretation at a location sets up a context leading to preferred interpretations at neighboring locations and so on. This process of 'growing' fragments of consistent interpretations is actually not very different from the way the Gibbs Sampler algorithm works in the context of simulated annealing. The literature on eye movements is vaguely supportive of this way of looking at things—a recent reference is (Kawabata 86).

Undoubtedly the proof of the pudding is in the eating and final judgement on this approach must be deferred till experimental results have been demonstrated.

VIEWPOINT INVARIANT PRIMITIVES AS A BASIS FOR VISUAL OBJECT RECOGNITION

Irving Biederman

University of Minnesota

ABSTRACT

One way to achieve real-time 3D object recognition from a 2D image is to posit an intermediate representation consisting of viewpoint invariant volumetric primitives. In Biederman's (1987) Recognition-by-Components (RBC) account of object recognition the primitives, called *geons*, are activated by categorical contrasts of viewpoint-invariant properties (VIPs) of image edges and vertices, such as straight vs. curved, parallel vs. non-parallel, and vertex type. The geons are robust to noise in that they can be identified even when portions of their edges and vertices are occluded or deleted. Objects are modeled as an arrangement of the geons. Research on two problems will be described: a) Evidence for intermediate (e.g., geon) representations, and b) a connectionist implementation of RBC (with John Hummel).

Evidence for intermediate representations

If the primitives are activated by image features (namely edges and vertices), why posit an intermediate primitive? Why not just represent an object as an arrangement of edges and vertices? One benefit of an intermediate primitive is noise and viewpoint invariance: the same geon can be activated even though different edges and vertices may be disrupted by noise or self occluded by variations in viewpoint. Several priming experiments with contour-deleted images provide evidence that this potential computational benefit may in fact be realized in real-time object recognition by humans.

Subjects first viewed a series of pictures of common objects, each with half its contour deleted by removing every other edge and vertex from each of its parts (geons). The subjects then identified briefly presented pictures of the same objects. The pictures were either identical to the original images or the complement to the original image. The complement had the remaining half of the edges and vertices. Performance (naming reaction times and error rates) was equivalent in the two conditions. This result suggests that the representation of an object is not the edges and vertices that are explicit in the image, but must be modeled as a more global representation, such as an arrangement of geons. Performance with the identical and complementary images was considerably superior to a condition in which the object model differed from the one viewed on the priming trial was shown. For example, in the latter condition, a grand piano might have been shown on the first trial and an upright piano on the second. This last result indicates that a substantial priming effect with the identical and complementary images is not merely due to repetition of the name or concept of the object or general practice on the task.

A Connectionist Implementation of RBC (w. John Hummel)

The model is designed to take as input a line drawing image of an object and, as output, activate a distributed representation of the geons and relations that make up an object model. Current work has focussed on the recognition of geons and the development of distributed representations for VIPS in hidden layers.

The model is a seven layer connectionist network. The model's input layer consists of 271 clusters of cells distributed in an hexagonal lattice over the model's visual field

(approximately the central 4 degrees of foveal vision). Each of these clusters contains six cells sensitive to edges within a particular range of orientations. All cells within an input cluster respond to edges within the same local region of the visual field. Hence, the input to the model is a representation of the edges in a visual image, coarsely coded with respect to orientation and location in the visual field.

The model's second layer is also arranged into 271 clusters of cells; each cluster in layer 2 corresponds to a single cluster in layer 1. The cells in these clusters respond to the VIPs posited by RBC. Thus, the cells in layer 2 respond to the vertices and axes of parallelism and symmetry defined by the configuration of edges in layer 1. These VIPs are detected independently for each location in the visual field and for each orientation in which they can occur. This type of representation is termed an enumerated representation, as the cells responding to the VIPs are enumerated over all locations and orientations in the visual field.

The model's third layer supports a semi-invariant representation of the VIPs detected in the second layer. A fully invariant representation of some feature (say, an arrow vertex) would be a representation that became active whenever that feature appeared in the model's visual field, regardless of the feature's scale, location or orientation. The representation in layer 3 is termed semi-invariant because it is invariant with location and scale but not with orientation. This layer also contains a group of cells that calculate the spatial relations among the VIPs represented in the second layer. Thus, the representation of the VIPs' orientations is absolute in this layer, but locations are represented relativistically.

The model's fourth layer uses the representation at the third layer to derive a semi-invariant representation of the geons in the image. As with the representation in the layer below, this representation is invariant with scale and position, but is enumerated over different values of orientation. In the fifth layer, orientation, too, is separated from the representation of the geon's other properties (such as type, scale, aspect ratio, etc.), and geons are represented in a fully invariant manner. The fifth layer also computes the spatial relations among the geons represented there. The resulting representation is invariant and explicitly expresses the relations among the geons in the original image. This representation is therefore suitable as a basis for invariant object recognition.

Although invariant object recognition could be performed directly from the representation in layer 5, many of the properties represented there will occur in conjunctions that may be germane to the identification of more than one object. For example, the representation of BRICK_UNDER_X may be used in identifying both an automobile (where X is a wedge) and a personal computer (where X is another brick). Therefore, the cells in layer 6 will be allowed to learn to respond to different combinations of such properties. In combination with an appropriate method for variable binding, this intervening representation will: (1) facilitate generalization to novel objects, (2) make more efficient use of representational resources, and (3) facilitate proper use of feature conjunctions (rather than simple feature lists). Naturally, the model's seventh layer is used to represent complete objects.

References

- Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94, 115-147.
- Hummel, J. E., Biederman, I., Gerhardstein, P. C., & Hilton, H. J. (1988). From image edges to geons: A connectionist approach. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.) *Proceedings of the 1988 Connectionist Models Summer School*. Pp. 462-471. San Mateo, CA: Morgan Kaufmann.

ABSTRACT: Representation and the Scene Modeling Problem

By: M.A. Fischler
SRI International
Menlo Park, Calif. 94025

The dominant paradigm in Artificial Intelligence in general, and in Machine Vision in particular, is largely based on the use of explicit models of discrete semantic objects, as a way of describing the world. Vision is considered to be the process in which sensor-derived signals are sequentially transformed into more organized and abstract descriptions of some viewed scene. The goal of machine vision research is, in the context of the above "signals-to-symbols" paradigm, to discover (or invent) an ordered set of representations* and algorithmic processes for transforming information contained in one representation into that of the next. For example, one specific set of such representations might involve first describing the image in terms of its variation in intensity, color, and texture, followed by the organization of the image into lines, edges, and regions. Next, information in the image is used to derive a description of the scene geometry in terms of a collection of continuous surfaces. The surfaces are then associated with coherent 3-D objects; and finally, the objects are recognized as instances of named entities.

The signals-to-symbols paradigm gives rise to a number of problems:

- a) (The Partitioning & Frame Problems) In the process of successive abstraction, how can we decide what information can be thrown away as irrelevant and how can we partition the image into meaningful entities, before we have determined what objects are actually present in the scene.
- b) (The Indexing Problem) Even if we can successfully partition the image into meaningful regions, or the scene into coherent objects, how do we know which model(s) to invoke as relevant to further interpretation and identification, short of trying all of the models in our database.
- c) (The Representation/Modeling Problem) How can we describe to the machine, in a computationally useful way, what a complex natural object (say a particular bush) looks like, or how to recognize such an object in an image as being an instance of some general class (e.g., bushes). If we must describe every leaf, and how the leaves are arranged, and how they can move between sightings, then we have an impossible task.

There are a number of additional problems of equal importance, but more difficult to present informally. All of the above problems are still unsolved in constrained domains (such as in indoor --- man-made --- environments) and most of these problems appear to be intractable in the design of a general purpose vision system competent to interpret unconstrained views of the natural outdoor world.

In this talk, I will explore some ideas in the areas of representation and modeling that have the potential to solve, or at least bypass some of the signals-to-symbols problems. In particular, I will discuss isomorphic/analogic representations, monolithic and physically motivated models, and interpretation based on the principle of simplest description.

In an *isomorphic representation* (some of) the semantics of the application domain are inherent in the data structure of the representation: e.g., an image formed on a retina or in a photograph is an isomorphic representation of the scene it depicts --- the imaging and photographic processes adequately preserve geometric adjacency, color, texture, and brightness.

In a *monolithic model*, there is no requirement for semantically meaningful substructure. For example, a set of equations can be used to describe the relation between scene geometry and image geometry --- intermediate results in the solution of the equations typically have no physical meaning. On the other hand, each of the intermediate representations in the signals-to-symbols formalism must have a semantic interpretation.

In a situation where explicit models do not exist for each of the objects of interest which might occur in a scene, there must be some generic basis for choosing one scene description over another. This can be accomplished in a principled way, for example, by choosing the {it simplest} of a set of competing descriptions.

The mechanism's biological systems use to convert visual signals into useful information is still a mystery we have barely begun to penetrate. It would be very rewarding if our attempt to build machine vision systems could make a contribution to this problem --- it still remains to be seen if the mechanisms we currently understand are indeed relevant.

*A representation is a data structure capable of encoding the information in a set of models.

Object Recognition from Range Images

Ramesh Jain

Artificial Intelligence Laboratory
Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109

In most applications of object recognition, a 3-dimensional object must be recognized from its 2-dimensional projections, which may be only partially visible in an image. Recognition of objects in images using exhaustive matching is a computationally hopeless task. The use of object features simplifies the recognition task. However, the task of determining effective features is usually that of the designer. It is desirable that the features used for recognition of objects be determined by the system using information available in CAD or other databases that contain models of all objects. In general, however, surprisingly little use has been made of such model databases in machine vision.

In recent years digitized range data has become available and the quality of this data has been steadily improving. Not only are depth relationships in range image regions explicit, the three-dimensional *shape* of the regions approximates the three-dimensional shape of the corresponding object surfaces in the field of view. Therefore, due to the explicitness of the information, the process of recognizing objects by their shape in range images *should* be less difficult than in intensity images.

We are developing techniques for using model databases for 3-D object recognition and inspection. Though currently we are using images acquired using laser range finders, our emphasis in this project is on developing techniques that will be applicable to intensity images also. Our focus is on object recognition; in most cases once the object is recognized, the location and orientation information can be easily recovered. We will

develop techniques for recovering robust features from images and use them in object recognition using the Feature Indexed Parallel Distributed Processing(FIPDP) network. We are also developing an algorithm that will generate aspect graphs for curved objects using their models in a CAD database, and help in automatic generation of a FIPDP network for recognition of objects.

PERCEPTION AND PERCEIVERS

A. Jepson and W. Richards

SUMMARY

Perception is our window to the world. Yet its essence remains somewhat elusive. For example, we have no formal definition of a perception. By formal we mean a definition precise enough to be captured by a computer program. Here, we offer such a definition. captured by a computer program. Here, we offer such a definition.

Intuitively, one might regard a perception as a successful match to an internal model. Usually this is what is implied by "shape perception", "scene perception", "motion perception", etc. For example, when we inspect a Julesz random-dot stereogram, the perception begins when a surface starts to coalesce from the random noise field and finishes with a particular shape sitting in 3D. Note that in this case, there need be no a priori, well defined model. The "shape" could just as well have been a snow storm as a flat planar surface. However, once a definite, stable shape is seen, we consider the perceptual act to have taken place, regardless how many other possible interpretations may be realizable. What conditions must be satisfied for us to elevate an interpretation of our sensory data to a perception, as opposed to mere sensations?

Our theory of "a perception" requires an internal conceptualization of our world. We assume that this conceptualization includes a categorization of events in the world, and a language for representing and reasoning about these events. In addition, the conceptualization provides knowledge and fallible beliefs both about the world and about how to organize the sensory data into a plausible world model. These intuitive notions are made precise in a formal system which we call a "perceiver". A "plausible" world model is defined to be a consistent interpretation of the image which satisfies a maximal set of beliefs. This maximal state can be computed using default logic. The perception, then, is taken to be the set of categories and relations between categories that are believed to be true in this maximal state. To illustrate these ideas we present a simple perceiver which uses several key elements of our formalization to compute a perception of an "image".

Parallel processing of form, color, motion, and depth in the visual system

Margaret Livingstone

Even though intuition suggests that our vision can plausibly be subdivided into form, color, depth, movement and texture perception, and perhaps a few others, our perception of any scene usually seems well unified. Despite this apparent wholeness, studies David Hubel and I have been doing on anatomy, physiology, and human perception are converging toward the conclusion that the visual system is subdivided into several separate parts whose functions are quite distinct.

Anatomical and physiological studies in monkeys support this idea of functional divergence within the visual pathway. They reveal major anatomical subdivisions at the earliest peripheral stages in the visual system as well as segregation of function at the highest known cortical stages, but until recently there was little information about subdivisions in the intermediate levels, the first and second cortical visual areas.

The subdivisions differ markedly in four major ways--in color, quickness, acuity and contrast sensitivity--implying that they contribute to different aspects of vision. Exactly *what* aspects have become clearer recently, when new anatomical techniques made it possible for us to follow these subdivisions farther into the central nervous system, so we could then determine the higher level response selectivity of cells at later stages in each subdivision.

There are strong suggestions that these channels remain segregated through still higher levels in the brain. From lesion studies in monkeys Pohl, Ungerleider, and Mishkin at NIH have defined two functionally distinct divisions of visual association areas: the temporal-occipital region, necessary for learning to identify objects by their appearance, and the parieto-occipital region, needed for tasks involving the positions of objects, a distinction they refer to as *where* versus *what*. Occasionally people with strokes will suffer surprisingly specific visual losses--for example, loss of color discrimination without impairment of form perception, loss of motion perception without loss of color or form perception, or loss of face recognition with preservation of the ability to recognize most other categories of objects as well as the ability to see color and depth, suggesting that in humans as well the visual pathway is functionally subdivided at a fairly gross level.

Our most recent research has been aimed at asking whether the differences seen at the early stages of these subdivisions can be detected in human visual perception by comparing the color, temporal, spatial and contrast sensitivities of different visual functions. Many of these questions have, not surprisingly, already been asked, and the answers are strikingly consistent with the anatomy and physiology. For several decades psychologists have accumulated evidence for two channels in human vision, one chromatic and the other achromatic, by showing that different tasks can have very different sensitivities to color and brightness contrast. Now that more is known about the electrophysiology and the anatomy of the subdivisions within the primate visual system, we can begin to try to correlate the perceptual observations with these subdivisions.

Electrophysiological studies suggest that one system, the magno system, is responsible for carrying information about movement and depth. We are extending our ideas about the possible functions of this system with perceptual studies, and suspect that the magno system may have the broader function of determining the spatial organization of elements in any visual scene. Magnocellular functions may include deciding which visual elements, such as edges and discontinuities, belong to and define individual objects in the scene, as well as determining the overall three-dimensional organization of the scene and the positions in space and movement of objects.

The other system, the parvo system, seems to be important for analyzing the scene in much greater and more leisurely detail and is particularly suited for the analysis and correlation of many kinds of visual properties and details, especially of static objects. These postulated functions would be consistent with the evolutionary relationship of the two systems: the magno system seems to be more primitive than the parvo system and is possibly homologous to the entire visual system of non-primate mammals. If so, it should not be surprising that magno system is capable of what seem to be the essential functions of vision for an animal that needs to use vision to navigate in its environment and to catch prey or avoid predators. The parvo system, which is well developed only in primates, seems to have added the ability to scrutinize in much more detail the shape, color, and surface properties of objects, enabling it to correlate and assign multiple visual attributes to a single object.

Such a functional segregation of visual information processing has important implications for people working in other visual fields, such as art, design, camouflage, radiology, radar, and design of artificial visual systems, and most recently I have been exploring these implications. For example, since the Gestalt of an image is totally dependent on lightness-contrast information, this means that lines that are meant to draw the eye or indicate shape, in a garment, an advertisement, or a picture, must have brightness contrast; color-contrast alone is inadequate. Moreover, since the movement-sensitive magno system is much more sensitive to low-contrast, then people interested in detecting low-contrast images, such as in radiology, radar, or in trying to see objects at night or in fog, should optimize the activity of the magno system by introducing movement, either of the object or of the observer.

CORTICAL PATHWAYS FOR THE ANALYSIS OF FORM, SPACE, AND MOTION:

THREE STREAMS OF VISUAL PROCESSING

Leslie G. Ungerleider

Laboratory of Neuropsychology, National Institute of Mental Health

Bethesda, Maryland 20892

We have previously proposed (Ungerleider and Mishkin, 1982) that striate cortex in the macaque is the source of two diverging cortical pathways: one, an occipitotemporal pathway, which enables the visual recognition of objects; the other, an occipitoparietal pathway, which mediates the appreciation of the spatial relationships among objects as well as the visual guidance of movement towards objects in space. The original behavioral evidence which led us to this proposal was the finding that monkeys with inferior temporal lesions are severely impaired on object recognition tasks but not on visuospatial tasks, whereas monkeys with posterior parietal lesions are impaired on visuospatial tasks but not on object recognition tasks.

There is considerable evidence from humans with brain damage for such a dissociation of function between the temporal and parietal lobes as well, and recently we've obtained evidence for this distinction from PET studies done in collaboration with scientists at the National Institute of Aging (Haxby et al., 1988). In these studies, we measured cerebral blood flow using ^{15}O -labeled water in normal subjects performing match-to-sample visual processing tasks. Our object vision task involved face recognition in which the subject had to select the face that matched the sample despite changes in the shadowing or orientation of the faces. Our spatial vision task involved perceiving the location of a dot in a square that contained a single border, and, in this

case, the matching stimuli were rotated relative to the sample. The results showed that cortical areas consistently activated during the face recognition task were in the region of the occipitotemporal junction but not in parietal cortex; moreover, the activation was bilaterally symmetrical. By contrast, areas activated more by the spatial vision task than by the face recognition task were in the superior parietal lobule; interestingly, in this instance, the activation was greater in the right than in the left hemisphere. The results thus demonstrate the existence in humans, as in monkeys, of two distinct processing systems, although there may be cross-species differences in their anatomical locations.

To understand the anatomical circuitry underlying these functions, we've attempted to differentiate all of the areas that comprise visual cortex in the macaque, and to trace the flow of visual information through them step-wise from the primary visual cortex to the highest-order visual areas in the temporal and parietal lobes (for review, see Desimone and Ungerleider, 1989). We have found that the areas along the occipitotemporal pathway (V1, V2, V4, and areas TEO and TE within the inferior temporal cortex) appear to be organized primarily as a serial hierarchy in which each area processes both color and form information. By contrast, the areas along the occipitoparietal pathway (V1, MT, and MT's multiple projection zones in parietal cortex) process the direction of stimulus motion and probably the relative spatial locations of stimuli. We have followed the projections of two of the areas to which MT projects within the superior temporal sulcus (Boussaoud et al., 1987), and have found that these two areas, MST and FST, have connections with widespread regions of the posterior parietal cortex, consistent with their role in visuospatial function. In addition, however, these areas also project extensively to portions of both the dorsal bank and floor of the anterior

superior temporal sulcus, including the superior temporal polysensory area (STP). Because MST and FST are both characterized by a high proportion of directionally sensitive cells and many cells in STP respond to complex stimulus motion, such as rotation and optical flow, the results suggest a third stream of visual processing, one which is concerned with motion analysis and extends from V1 to MT and then forward into the superior temporal sulcus.

Thus, we propose that there may be three cortical streams of visual processing: an occipitotemporal stream for color and form, an occipitoparietal stream for spatial location, and an occipito-superior temporal stream for motion.

Components of High-Level Vision

Stephen M. Kosslyn

Vision can be disrupted in a wide variety of ways following brain damage. We assume that these qualitatively distinct types of impairments reflect in part the structure of the underlying processing mechanisms. The goal of this project is to understand impairments of visual object identification as a consequence of damage to individual "processing subsystems" and their interconnections. A processing subsystem, as here conceived, corresponds to a set of neurons that work together to accomplish part of an information processing task. Each subsystem is characterized in terms of its input, operation, and output, in the context of specific assumptions about the nature of the environment (following Marr). Because we are interested in understanding the effects of brain damage on information processing, we attempt to specify what is accomplished by (rather large) portions of neural tissue. We make no commitments regarding the algorithms that accomplish specific input/output mappings, but only to the claim that certain classes of such operations are performed.

The theory of processing subsystems that guides the present research was developed in light of three kinds of information. First, we considered the most fundamental behavioral abilities of the object recognition system as a whole. Without knowing what the normal system can do, we are in no position to understand the underlying mechanisms (which when damaged impair function). The basic abilities we consider fall into three classes: First, objects can be identified when seen from different points of view (and hence their images project different shapes, different sizes, or at different positions in the field). Second, objects can be identified even when they assume atypical shapes (as occurs when optional parts are added or deleted, such as arms on a chair; when the shapes of parts change, as occurs for backs of chairs; and when the spatial relations among parts change, as when a dog is jumping versus curled up sleeping). Third, objects can be identified even when only partial information is available (as occurs when an object is partially occluded).

Once we have characterized the fundamental behavioral abilities of the system, we are in a position to begin formulating a theory of the underlying mechanisms that produce this behavior. We do not attempt to account for the observed behavioral abilities in detail; rather we use this characterization to ensure that our theory is not in principle incorrect and to sketch out very broad classes of mechanisms that in principle could produce the observed abilities. The theory is also constrained in part by facts about the neuroanatomy and neurophysiology of primate vision. In particular, the theory rests on the general observations that: a) there are numerous areas in the primate brain concerned with vision (approximately 30 in the macaque brain, at last count); b) these areas are hierarchically organized; c) areas receiving afferent projections from other areas in turn have efferent projections (of comparable size) back to those areas; d) the visual areas are organized into

functional streams; and, e) areas in the different streams may be structured differently, and neurons in these areas often have different physiological properties; we have taken such characteristics of specific areas as hints about possible functions of the areas. (This information was drawn from the work of Allman, Desimone, Gross, Livingstone & Hubel, Maunsell, Mishkin, Ungerleider, Van Essen, and others.) Many of these starting points (for us) will be discussed in detail by other presentations at this conference.

Finally, we attempt to characterize (at a coarse level) the processing subsystems that could allow a system with this neural substrate to produce the observed behavioral properties. The theory is developed by performing analyses of the computations that appear to be necessary for such a system to produce such behavior (deriving poor-man's versions of what Marr called the "theory of the computation"). These hypotheses are tested by observing the behavior of both the normal system (e.g., measuring the time to identify objects under various circumstances) and the damaged system (e.g., dysfunctions in object identification following stroke).

The analyses offered here result in an hypothesized decomposition of the high-level visual system into five major components:

A visual buffer. Retinotopically mapped areas in the occipital lobe post-V1 and V2 are treated as a single functional structure. This structure represents input at multiple scales of resolution. An "attention window" selects some region, at some resolution, within this structure for further processing (cf. Moran and Desimone, Treisman).

Object properties encoding. The contents of the attention window are sent to nontopographically mapped areas in the inferior temporal lobe, which are posited to extract "nonaccidental" properties (in the sense of Lowe and Biederman) and use them to access stored visual memories. These memories represent shape, color, and texture (cf. Gross, Mishkin, Desimone, et al.).

Spatial properties encoding. At the same time that the contents of the attention window are sent to the inferior temporal lobe, they also are sent to the spatial properties encoding subsystems in the parietal lobe. These subsystems are posited to extract (and store memories of) location, size, and orientation. (Ungerleider, Mishkin, Andersen, Maunsell, Van Essen and others have described this division of processing.)

Multimodal associative memory. The two classes of encoding subsystems in turn provide input to a multimodal associative memory, which involves temporal lobe and frontal lobe structures. This memory contains pointers that cross-index information in different sensory modalities. Object representations in this memory include associated names, salient properties, functions, and contexts. (These claims are informed by the work of Goldman-Rakic, Gross, and others.)

Top-down hypothesis testing. If input matches stored information poorly when an object is first seen, additional encoding cycles are required. Salient properties of the most likely object are accessed in associative memory, and then used to guide attention to the location at which a distinctive object property should be found if the hypothesis is correct (cf. Gregory, Neisser). These processes rely on frontal lobe structures (area 8, dorsolateral prefrontal) and parietal lobe structures (cf. Luria).

Each of these subsystems in turn is decomposed into several more specialized hypothesized processing subsystems. In several cases we are able to show that some of the subsystems are more effective in one of the two cerebral hemispheres. For example, we posit that the spatial properties encoding subsystems include one that encodes "categorical" spatial relations (e.g., "left of," "above," "connected to") and one that encodes actual metric locations; a series of converging experiments demonstrated that the categorical spatial relations encoding subsystem is more effective in the left cerebral hemisphere, whereas the metric spatial relations encoding subsystem is more effective in the right cerebral hemisphere. Thus, we have evidence not only for the distinction between the two, but also for their neural realization. A computer simulation model has been implemented in which one can simulate damage to these individual subsystems and their interconnections. This simulation produces a host of specific deficits in object identification that mirror those previously reported in humans following focal brain damage. In addition, the simulation makes predictions about previously unreported dysfunctions.

DEPTH CUES AND THE COMPUTATION OF REAL DISTANCE: THE CALIBRATION OF BALLISTIC MOVEMENTS

Melvyn A. Goodale

University of Western Ontario
London, Ontario, Canada

How the brain transforms the two-dimensional retinal image into a three-dimensional representation of the external world has been a problem that has occupied the attention of philosophers and scientists for hundreds of years. Nevertheless, much of the research and thinking on this problem, even in modern times, has "lost sight" of the evolutionary context in which visual systems developed. As I have argued elsewhere (Goodale 1988), vision evolved, not to provide the organism with a unified percept of the world in which it lives, but to control the movements that the animal makes in that world. Natural selection operates at the level of overt behavior. It is not interested in how well an animal "sees" the world in which predators and prey can be found, but only in how well the animal avoids the predators and catches the prey. Indeed, an argument can be made that it was not a visual system that evolved but a visuomotor one. If this is the case, then the functional architecture of such a system can be fully understood only by studying its motor outputs as well as its sensory inputs.

The failure to appreciate the important contribution that visual information makes to the control of motor outputs is particularly evident in the study of "depth vision". Most investigations of depth vision in mammals have required animals to discriminate between two stimuli either on the basis of their relative depth or on the basis of a difference in form that is revealed only by a single depth cue such as retinal disparity. Only rarely have animals been required to estimate the actual distance of the stimuli that are presented.

In the real world, of course, depth vision is often used to calibrate motor output as a function of the exact distance of an object or surface. This is particularly true for ballistic movements where the animal has little opportunity to adjust the trajectory of the movement as it unfolds. In our laboratory, we have been studying one kind of visually controlled ballistic movement -- jumping in the Mongolian gerbil (*Meriones unguiculatus*). All the gerbil has to do in our situation is jump from one platform to another to obtain a shelled sunflower seed. The performance of the gerbil is recorded on videotape using strobe-shutter cameras. Although the distance between the two platforms varies randomly from trial to trial, the gerbils are very accurate in matching the amplitude of their jump to the size of the gap. We have demonstrated that the vertical translation movements (or head bobs) that the gerbil often makes before jumping are almost certainly used to generate

motion cues that are then used to calibrate the amplitude of the jump. In a number of experiments (Ellard, Goodale and Timney, 1984; Ellard, Goodale, MacLaren Scorfield, and Lawrence, 1986), we have found that there was not only a significant correlation between the overall accuracy of an animal's jumps and the incidence of head bobs preceding those jumps, but limiting the availability of other distance cues, such as loom and stereopsis, increased the likelihood that an animal would bob its head before jumping. While the use of retinal motion ("peering") has been demonstrated in the desert locust, *Schistocerca gregaria*, by Wallace (1959) and later by Collett (1978), our work represents the first clear demonstration in a mammal (other than the human) that the depth information generated by translational movements can be used to calibrate motor output. The gerbil apparently calibrates its jump by computing the relationship between the velocity (or displacement) of the retinal image of the platform, the velocity (or displacement) of the eyes in the orbit as they fixate the platform, and the velocity (or amplitude) of the vertical head movement. In other words, the retinal contribution to the distance equation appears to be derived entirely from the motion of the image of the landing platform and not from the parallax generated by the relative motion of the image of the platform with respect to the image of the background. Indeed, while the latter cue provides relative distance information in the absence of information about the velocity (or amplitude) of the translation movements that generate the parallax, it will not on its own provide any information about the absolute distance of an object.

In later experiments, we showed that the size of the retinal image size of the landing surface, which varies inversely with distance, acts as a kind of range finder for the head-bob system. Generating motion cues by bobbing one's head has costs as well as benefits in the real world of the gerbil. While a few seconds of head bobbing would certainly provide the animal with more accurate distance information for calibrating a jump, those few seconds could also provide a predator with an opportunity to make an easy catch. For this reason, it would be useful for the gerbil to have a rough idea of how far away the landing place is, so that it can decide whether or not it should pause and bring the more sensitive motion system to bear. The margin for error might be small and if the intended target is some distance away, then the distance information provided by motion cues might make the difference between a successful and an unsuccessful jump. Retinal image size appears to supply this rough estimate. In addition, information derived from retinal image size contributes directly to the calibration of jump amplitude.

It is important to remember, however, that in its natural environment a gerbil might sometimes be in unfamiliar terrain and be required to estimate the distance of unfamiliar objects. In this situation, the size of an object's retinal image would provide almost no information about the object's distance. Other distance cues,

however, such as retinal motion, would continue to be useful since they are largely independent of the gerbil's familiarity with the object. In light of this, we were not surprised to observe that gerbils trained to jump to objects that vary in size from trial to trial make more head bobs than gerbils trained to jump to a familiar landing surface.

Recent work in our laboratory has shown that while striate cortex and its cortical elaboration may be involved in the computation of absolute distance on the basis of the motion of the retinal image in the gerbil, this projection system is not essential for the computation of distance on the basis of retinal image size. At the same time, size constancy, which also involves the use of retinal image size, does appear to depend on mechanisms that derive their input from striate cortex. This dissociation between visuomotor mechanisms on the one hand and more "perceptual" systems on the other characterizes visual systems in a number of different mammals, including humans. We have shown, for example, that human subjects will fail to perceive changes in the position of a target even though those changes will elicit large modifications in the trajectory of a limb movement directed at that target (Goodale, Pelisson, and Prablanc, 1986). Even more striking dissociations between perceptual report and the control of motor output have been observed in patients with cortical lesions (Goodale, 1988). These and other results suggest that until we are prepared to study the outputs of the mammalian visual system as carefully as we study its inputs, our understanding of its underlying neural architecture and functional organization will remain quite incomplete.

REFERENCES

- Collett, T. (1978). Peering: A locust behaviour pattern for obtaining motion parallax information. J. exp. Biol., 76, 237-241.
- Ellard, C. G., Goodale, M. A., and Timney, B. (1984). Distance estimation in the Mongolian Gerbil: The role of dynamic depth cues. Behav. Brain Res., 14, 29-39.
- Ellard, C. G., Goodale, M. A., MacLaren Scorfield, D. M., and Lawrence, C. (1986). Visual cortical lesions abolish the use of motion parallax in the Mongolian gerbil. Exp. Brain Res., 64, 599
- Goodale, M. A. (1988). Modularity in visuomotor control: from input to output. In Computational Approaches to Human Vision: an Interdisciplinary Perspective (Edited by Pylyshyn, Z.) pp. 262-285. Norwood, New Jersey: Ablex.
- Goodale, M. A., Pelisson, D., and Prablanc, C. (1986). Large adjustments in visually guided reaching do not depend on vision of the hand or perception of target displacement. Nature, 320, 748-750.
- Wallace, G. K. (1959). Visual scanning in the desert locust *Schistocerca gregaria* forskal. J. Exp. Biol., 36, 512-525.

Constraints Imposed by Occlusion and Image-Segmentation

**V. S. Ramachandran
Psychology Department
University of California, San Diego
La Jolla, CA 92093**

The first step toward understanding any complex information processing system is to clearly identify the problems it was designed to solve. The "computational" approach to vision has been extremely useful in this regard because it allows a much more rigorous formulation of perceptual problems (Poggio et al., 1985; Ullman, 1979) than what would be possible with psychophysics or physiology alone.

1. The constraints imposed by the environment (natural constraints) reduce the computational burden on the visual system but they do not impose a unique solution to perceptual problems. There are often many different ways of solving a problem theoretically and the only way to distinguish between them is to do old-fashioned psychophysics and neuroanatomy.
2. The central dogma of computational vision has been that the strategies used by any complex information processing system can be understood independent of hardware implementation. Contrary to this we would argue that biological vision is strongly constrained by the actual neural machinery that mediates it (e.g. see Livingstone & Hubel, 1987; Ramachandran & Gregory, 1978; Cavanagh, et al., 1983). There may be certain things that neurons simply cannot do and this automatically eliminates a wide range of theoretically plausible solutions.
3. For any given perceptual problem biological systems often seem to use multiple parallel mechanisms which exploit multiple constraints. Why use multiple mechanisms when a single one will suffice on computational grounds? There are at least two reasons. First, by using multiple strategies for any one problem, the system can get away with each of them being relatively crude and, therefore, easy to implement in real neural hardware (a bit like two drunks who can't walk individually but can manage to do so by leaning on each other for support!). Second, the simultaneous use of multiple parallel short-cuts allows more rapid processing of images and a greater tolerance for noise than what would be possible with a single sophisticated algorithm. It is this remarkable tolerance for noisy (sometimes camouflaged) images that characterizes biological vision and sets it apart from machine vision.

Motion correspondence

How does the visual system match successive "snapshots" of a moving object to generate an impression of smooth, continuous motion? Does it first extract 3-D shapes and outlines from the image and then proceed to match these or is motion correspondence based on a primitive point-to-point matching of luminance distribution? Our experiments suggest that the former strategy is used. Even relatively abstract stimulus features such as equiluminous texture-borders (Ramachandran, et al., 1973), illusory contours (Ramachandran, 1985) and shape-

from-shading (Ramachandran, 1988) can provide an input to the long-range motion system. Of course, moving objects usually differ from the background in terms of their surface reflectance, so under ordinary circumstances the visual system could rely entirely on luminance edges to achieve correspondence. But by using a variety of inputs (such as texture edges, and chromatic edges), the system is able to tolerate noisy images of the kind it would encounter in the natural world (e.g. a leopard moving against a screen of fluttering foliage). As alluded to earlier this is one of the major advantages of using multiple strategies for the same visual process.

But if correspondence is established mainly between "coarse" or "salient" features - what about the finer features in the image? When matching successive snapshots of a moving textured object (such as a leopard), how does the visual system know which spot goes with which? Our results suggest that the salient features are extracted and matched first and the unambiguous motion signal derived from them is blindly applied to all the finer image features - "motion-capture" (Ramachandran & Inada, 1985). By adopting this "short-cut" the visual system avoids the computational burden of having to keep track of all the individual spots.

Constraints imposed by occlusion

We have noted that "illusory" contours implied by occlusion can influence motion correspondence (Ramachandran, 1985). We will now consider two further examples.

In the first experiment an illusory square was made to jump left and right and exchange places with a single black spot which moved in the opposite direction. If retinal disparities were introduced so that the spot was stereoscopically in front of the plane implied by the discs (and illusory square) its left-right apparent motion could be seen vividly. On the other hand, if the spot was stereoscopically behind the plane of the illusory square its motion was "vetoed" or inhibited by the moving illusory square which now appeared to occlude it. These observations demonstrate a powerful interaction between occlusion, motion and stereopsis.

Our second stimulus was an ambiguous apparent motion display. Two spots were flashed on diagonally opposite corners of a square and then replaced by two spots appearing as the remaining two corners. The display was bistable and one could see either vertical or horizontal apparent motion. If five such displays were viewed simultaneously they become "synchronized" - i.e. the same motion direction was seen in all of them (Ramachandran and Anstis, 1986). Next, we used masking tape to occlude two spots in one of the five displays. Interestingly the spots in this display continued to move in synchrony with the other as though they were pairing with spots behind the occluder!

These experiments demonstrate that the solution to the correspondence problem powerfully constrained by occlusion. Similar effects can be demonstrated for the "aperture" problem (Ramachandran and Rogers-Ramachandran, 1989; see also Nakayama and Shimojo, 1989).

Stereopsis

AI researchers often use stereopsis as an example to illustrate that visual mechanisms are highly modular (Marr, 1981). A Julesz random-dot stereogram, for example, evokes a powerful sensation of depth even though it is completely devoid of other depth cues and contains no monocularly visible shapes or contours. One may be tempted to conclude, therefore, that stereopsis is a simple point-to-point matching process that does not interact significantly with other visual mechanisms.

We constructed stereograms out of "illusory contours" - contours that are invoked by the visual system to account for surprising gaps in the visual image. When a stereogram of this kind is superimposed on a regular grid of spots ("wallpaper") the spots corresponding to the illusory squares get captured or pulled forward so that they come to occupy the same stereoscopic depth plane as the illusory square (Ramachandran, 1986), even though the dots themselves do not convey any disparity information.

If the two eye's pictures were reversed to convey uncrossed disparities an illusory square was no longer

Binocular representation of the visual field in primate cortex

Eric L. Schwartz

Associate Professor of Brain Research, NYU Medical Center

Adjunct Associate Professor of Computer Science, Courant Institute of Mathematical Sciences

The binocular visual field is initially represented in striate cortex in the form of two full topographic maps (left and right eye), which are interlaced into the strip-like pattern of the ocular dominance column system. In order to model this architecture, several experimental and computational questions are prominent:

- 1.) What is a correct model for monocular topography in macaque V1?
Recent computer reconstructed 2DG mappings from our lab will be presented which provide an accurate model of macaque V1 topography, and which also clarify some of the confusion which has been raised during the past several years concerning the details of macaque V1 monocular topography.
- 2.) What techniques are needed to model the V1 monocular topographic map?
Examples of numerical conformal mapping and texture mapping which provide a wide field full acuity V1 map of natural scenes will be shown.
- 2.) What is a correct model for the ocular dominance column pattern of V1?
Computer flattened reconstructions of the full macaque ocular dominance column pattern will be shown, along with a parametric analysis of this system.
- 3.) How can the joint presence of two full maps, interlaced as columns, be reconciled with the simpler notion of a "topographic map"?
An algorithm and the associated image warping which allows the mapping of natural binocular scenes onto the joint columnar/topographic map of layer IV of macaque V1 will be discussed.
- 4.) What is the computational utility of representing multiple related visual maps in the form of "columns"?
A recent model of binocular stereo segmentation on a columnar architecture (Yeshurun and Schwartz, PAMI,6/89) will be presented. This model, based on the properties of the two-dimensional cepstral filter, is a "one-pass" non-linear spatial filter which has very good computational properties on conventional computer architectures, and has some interesting relationships to the heuristic "trade-offs" characteristic of human stereo vision.

seen and a completely new percept emerged. Four black "holes" were seen (corresponding to the discs) and through these holes one could see the four corners of a partially occluded square. The spots on the wallpaper were now "captured" by the corners alone rather than by the whole square (Ramachandran, 1986). These results imply, contrary to earlier claims, that the stereoscopic matching process is profoundly influenced by image segmentation and by occlusion (Ramachandran, 1986).

Shape-from-shading

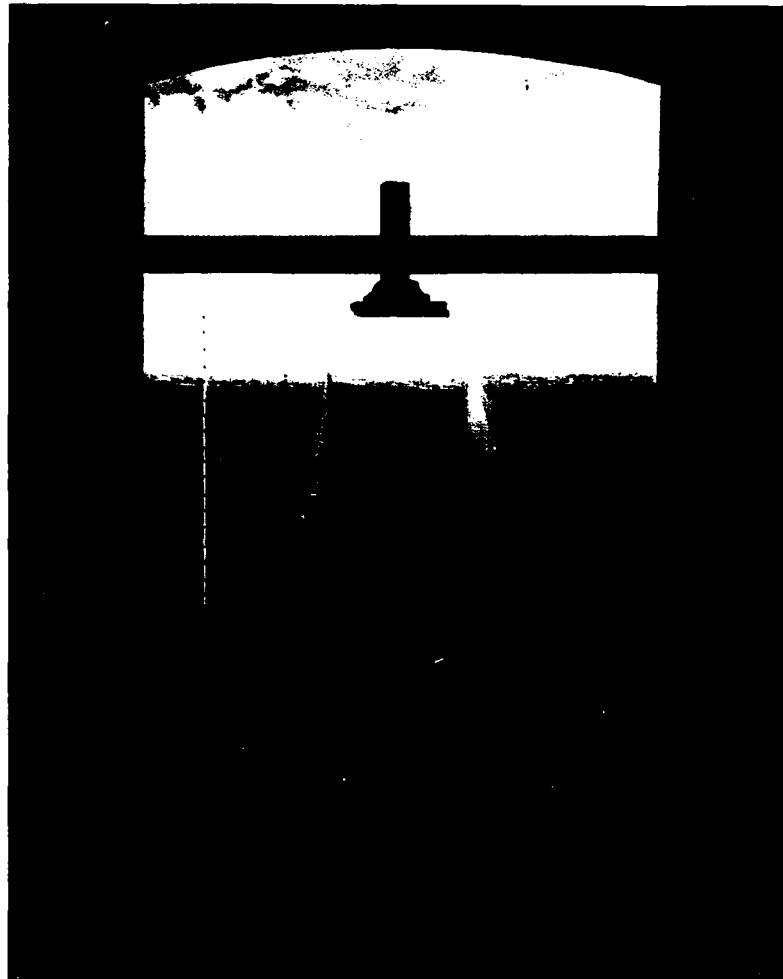
Our results suggest that the recovery of shape-from-shading is based on the combined use of occlusion boundaries and luminance gradients and not on detailed measurement of luminance variations. For example, a kanizsa-type "illusory circle" superimposed on a static one-dimensional luminance-ramp creates the compelling impression of an illusory sphere. Even though there is no sudden change in luminance across the border of the sphere one perceives a sphere because the simultaneous presence of the occlusion border and the luminance-ramp mutually reinforce that interpretation (Ramachandran, 1988B).

We find once 3-D shapes have been computed by the combined use of shading and occlusion they can be used as tokens for a variety of other visual capacities such as apparent motion, symmetry and perceptual grouping.

References

- Cavanagh, P., Tyler, C. & Favreau, O. (1983) *JOSA*, A1, 893-899.
- DeYoe, T. & Van Essen, D. (1988). *TINS*.
- Livingstone, M. & Hubel, D. H. (1987). *J. Neuroscience*, 7(11), 3416-3468.
- Marr, D. (1981). *Vision*, W. H. Freeman & Co., San Francisco.
- Nakayama, K. & Shimojo, S. (1989). *Vis. Res.*, in press.
- Poggio, T., Torre, V., & Koch, C. (1985). *Nature*, 317, 314-319.
- Ramachandran, V. S. (1985). *Perception*, 14, 97-103.
- Ramachandran, V. S. (1987). *Nature*, 138, 645-47.
- Ramachandran, V. S. & Cavanagh, P. (1987). *Vision Res.*, 27, 97-106.
- Ramachandran, V. S. (1988). *Scientific American*, 259, 76-83.
- Ramachandran, V. S., Rao, V. M. & Vidyasagar, T. (1973A). *Vision Research*, 13, 1399-1401.
- Ramachandran, V. S. & Inada, V. (1985). *Spatial Vision*, 1, 57-67.
- Ramachandran, V. S., Anstis, S. M. (1986). *Scientific American*, 254, 102-109.
- Ramachandran, V. S. & Gregory, R. L. (1978). *Nature*, 275, 55-56.
- Ramachandran, V. S. & Rogers-Ramachandran, D. (1989). *Soc. for Neuroscience Abstracts*.

VISION AND THREE- DIMENSIONAL REPRESENTATION



May 24 - 26, 1989

University of Minnesota

Sponsored by

Air Force Office of Scientific Research

College of Liberal Arts, University of Minnesota

Center for Research in Learning, Perception, and Cognition

■ ■ ■

VISION AND THREE-DIMENSIONAL REPRESENTATION

May 24-26, 1989 Electrical Engineering and Computer Science Building University of Minnesota Minneapolis, Minnesota

The appearance of the three-dimensional world from images projected on our two-dimensional retinas is immediate, effortless, and compelling. Despite the vigor of research in vision over the past two decades, questions remain about the nature of three-dimensional representations and the use of those representations for recognition and action. What information is gathered? How is it integrated and structured? How is the information used in higher level perceptual tasks? This conference will bring together 19 prominent speakers to address these questions from neurophysiological, psychological, and computational perspectives.

An art exhibit reflecting the theme of the conference will be held at the Coffman Gallery, Coffman Memorial Union, throughout May. For more information about the exhibit, contact Jay Barnes at 625-9523.



Cover photo: *Les Promenades d'Euclide, 1955, by Rene Magritte, The Minneapolis Institute of Arts*

PROGRAM

May 24

Electrical Engineering and Computer Science Building
Room 3-180

8:00 a.m. Opening

8:30

Development of Depth Perception
Albert Yonas, Institute of Child Development, University of Minnesota

Binocular Space Perception
John Foley, Department of Psychology, University of California, Santa Barbara

10:20 Break

10:30

The Reconstruction of Continuous Surfaces from Stereo Measurements and Monocular Inferences
Kent Stevens, Department of Computer Science, University of Oregon

Occlusion Constraints and the Encoding of Color, Form, Motion, and Depth
Ken Nakayama, Smith-Kettlewell Eye Institute, San Francisco

12:20

Lunch (on your own)

2:00 p.m.

Perception of 3-D Structure from Motion
James Todd, Department of Psychology, Brandeis University

Analyzing Visual Motion—Spatial Organization at Surface Boundaries
William B. Thompson, Department of Computer Science, University of Minnesota

3:30 Break

3:45

Affine Shape from Motion
J.J. Koenderink, Physics Laboratory, State University, Utrecht

4:50

Dinner (on your own)

7:00

A Perceptual Approach to Art: Comments on the Art Exhibition, Mississippi Room, Coffman Memorial Union
Bruce Goldstein, Department of Psychology, University of Pittsburgh

A reception and viewing of the art exhibition in the Coffman Gallery will follow.

May 25

8:00 a.m.

How 3-D Are We?
Patrick Cavanagh, Department of Psychology, University of Montreal

Representing Constraints for Inferring 3-D Scene Structure from Monocular Views

Jittendra Malik, Department of Computer Science, University of California, Berkeley

9:50 Break

10:00

Representation and the Scene Modeling Problem
Martin Fischler, SRI International, Menlo Park, California

3-D Recognition from Range Imagery
Ramesh Jain, Department of Electrical Engineering and Computer Science, University of Michigan

11:50

Lunch (on your own)

1:30 p.m.

What Must We Know to Recognize Something
David Lowe, Department of Computer Science, University of British Columbia

Viewpoint Invariant Primitives as a Basis for Human Object Recognition
Irving Biederman, Department of Psychology, University of Minnesota

3:20 Break

3:30

Separate Processing of Form, Color, Movement and Depth: Anatomy, Physiology, Art, and Illusion
Margaret Livingstone, Department of Neurophysiology, Harvard Medical School

Components of High-Level Vision
Stephen Kosslyn, Department of Psychology, Harvard University

5:20 Conclusion

May 26

8:00 a.m.

Depth Cues and Distance Estimation: The Calibration of Ballistic Movements
Melvin Goodale, Department of Psychology, University of Western Ontario

Cortical Pathways for the Analysis of Form, Space, and Motion: Three Streams of Visual Processing
Leslie G. Ungerleider, NIMH Laboratory of Neuropsychology, Bethesda, Maryland

9:30 Break

9:45

Binocular Representation of the Visual Field in Primate Cortex
Eric Schwartz, New York University Medical Center, Brain Research Laboratory

11:00 Closing

■ ■ ■

PROGRAM SPONSORS

Air Force Office of Scientific Research
College of Liberal Arts, University of Minnesota

Center for Research in Learning, Perception, and Cognition

In cooperation with the Departments of Computer Science, Electrical Engineering, Psychology, Studio Arts, the Institute of Child Development, and Professional Development and Conference Services, Continuing Education and Extension.

CONFERENCE PLANNING COMMITTEE

Martha Arterberry, Jay Barnes, Irving Biederman, Lisa Brienzo, Victor Caglioti, Lincoln Craton, Steven Cusulos, Jon Gottesman, Char Greenwald, John Hilton, Jonathan Marshall, Jo Nichols, Bruce Overmier, David Parish, and Liz Stuck

CONFERENCE CHAIRPERSONS

Lee Zimmerman, Electrical Engineering, (612) 625-8544
Gordon Legge, Department of Psychology, (612) 625-0846

REGISTRATION

The conference fee is \$30 (\$15 for current students). This fee includes program materials, refreshments, and Wednesday's reception. Conference enrollment is limited, so early registration is recommended. All registrations must be received by May 15. A refund, less a \$15 cancel-

lation fee, will be made if the registration is cancelled five working days prior to the conference. The University of Minnesota reserves the right to cancel the conference if necessary.

LOCATION/PARKING

The conference will be held in Room 3-180 Electrical Engineering and Computer Science Building, University of Minnesota, Minneapolis. Parking is available nearby in the Harvard Street Ramp, 216 Harvard Street S.E. A map indicating building and parking locations will be sent to registrants.

ACCOMMODATIONS

A block of rooms has been reserved at the Radisson University Hotel. Rates are \$68 (plus tax) for double or single occupancy. To make reservations, contact the hotel at (612)379-8888 and note the program title to obtain these special rates. Reservations must be made by April 9.

For further information, contact:

Program: Jo Nichols, Center for Research in Learning, Perception, and Cognition, (612) 625-9367

Registration: Char Greenwald, Professional Development and Conference Services, (612) 625-1520

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, religion, color, sex, national origin, handicap, age, veteran status, or sexual orientation.

Registration Form

54-38LB

Vision and Three-Dimensional Representation
University of Minnesota

— I enclose \$30 general registration.

— I enclose \$15 current student registration. Student I.D number:

— The above fee will be provided by the University of Minnesota.

Department budget number:

Name

Address

Telephone (business) (home)

Position

Affiliation

Please duplicate for additional registrations.

Please make check or money order payable to the University of Minnesota.

Mail to:

Registrar

Professional Development and Conference Services

University of Minnesota

338 Nolte Center

315 Pillsbury Drive S.E.

Minneapolis, MN 55455-0139

Registration should be received by May 15.